On Membership Inference Attacks to Generative Language Models across Language Domains

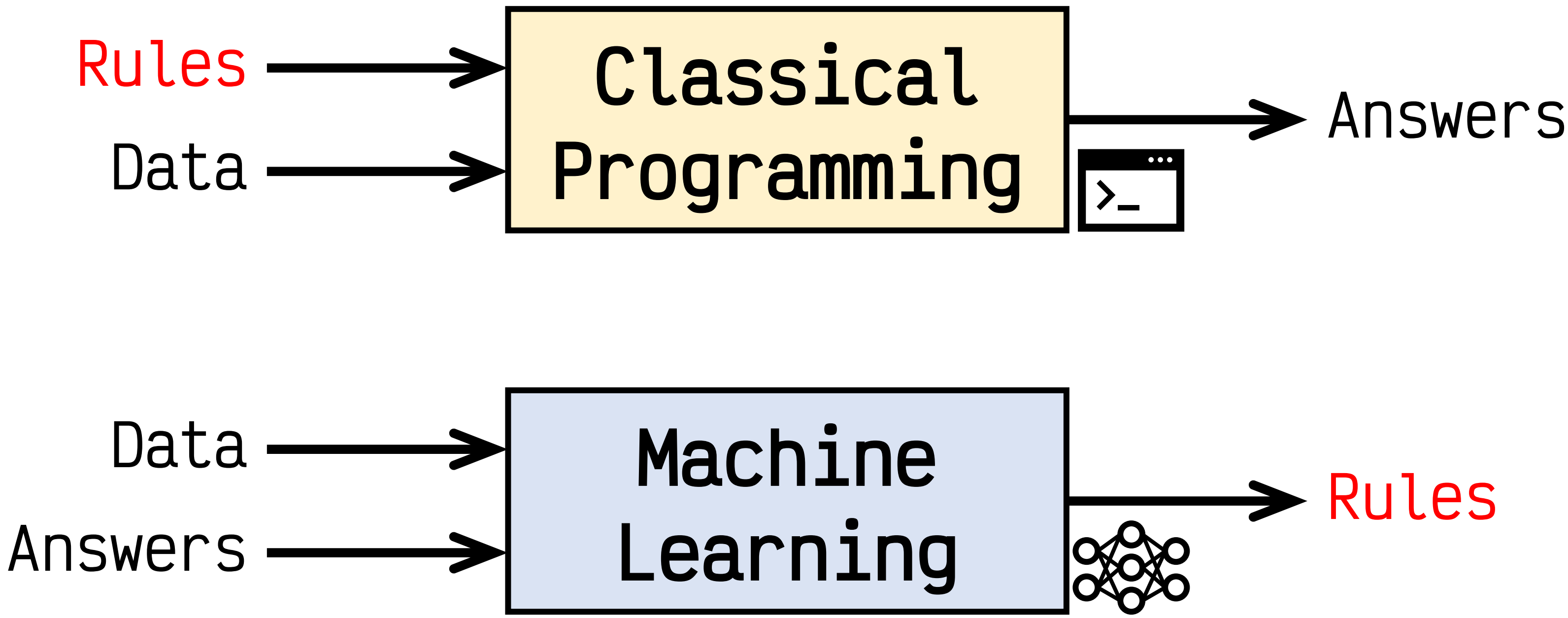
Myung Gyo Oh[*], Leo Hyun Park, Jaeuk Kim, Jaewoo Park, and Taekyoung Kwon[†]

Yonsei University

연세대학교 정보보호연구실
INFORMATION SECURITY LAB @ GSI
YONSEI UNIVERSITY, SEOUL, KOREA

Machine Learning

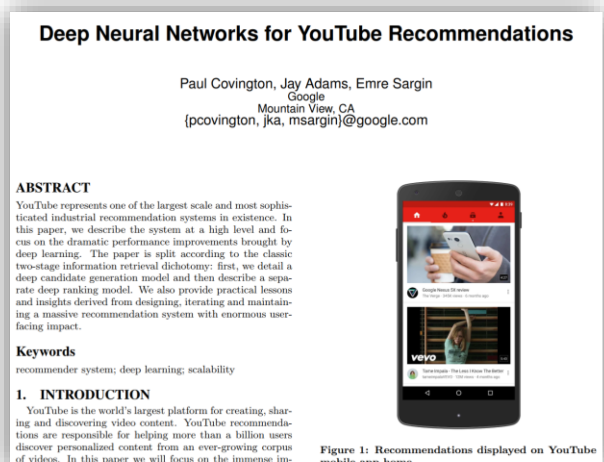
- Classical Programming vs. Machine Learning



https://kislayverma.com/programming/combining-rule-systems-and-machine-learning/ *(reconstruct)*

Deep Learning in Real World

- Object Detection
  - Autonomous driving (Tesla)
- Recommendation System
  - Netflix [Gomez-Uribe et al. (2015)]
  - YouTube [Covington et al. (2016)]



Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin
Google
Mountain View, CA
{pcovington, jka, msargin}@google.com

ABSTRACT

YouTube represents one of the largest scale and most sophisticated industrial recommendation systems in existence. In this paper, we describe the system at a high level and focus on the dramatic performance improvements brought by deep learning. The paper is split according to the classic two-stage information retrieval dichotomy: first, we detail a deep candidate generation model and then describe a separate deep ranking model. We also provide practical lessons and insights derived from designing, iterating and maintaining a massive recommendation system with enormous user-facing impact.

Keywords

recommender system; deep learning; scalability

1. INTRODUCTION

YouTube is the world's largest platform for creating, sharing and discovering video content. YouTube recommendations are responsible for helping more than a billion users discover personalized content from an ever-growing corpus of videos. In this paper we will focus on the immense im-

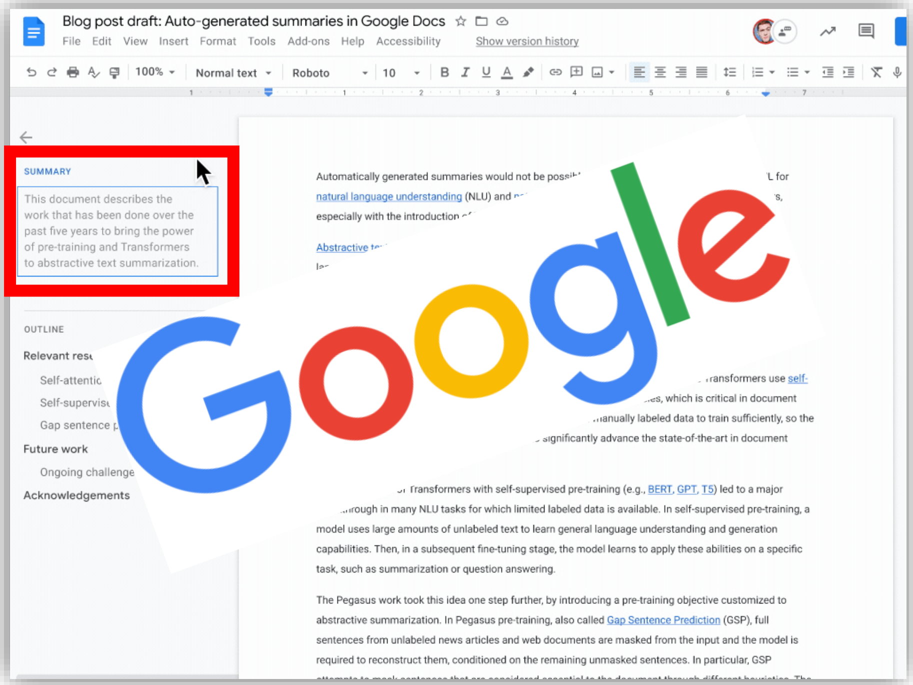
Figure 1: Recommendations displayed on YouTube mobile app home.

[Covington et al. (2016)]

https://www.tesla.com/AI

Deep Learning in Real World

- Generative Language Model (LM)
  - Abstractive summarization (Google PEGASUS) [Zhang et al. (2020)]
  - Coding suggestion (GitHub Copilot) [Chen et al. (2021)]

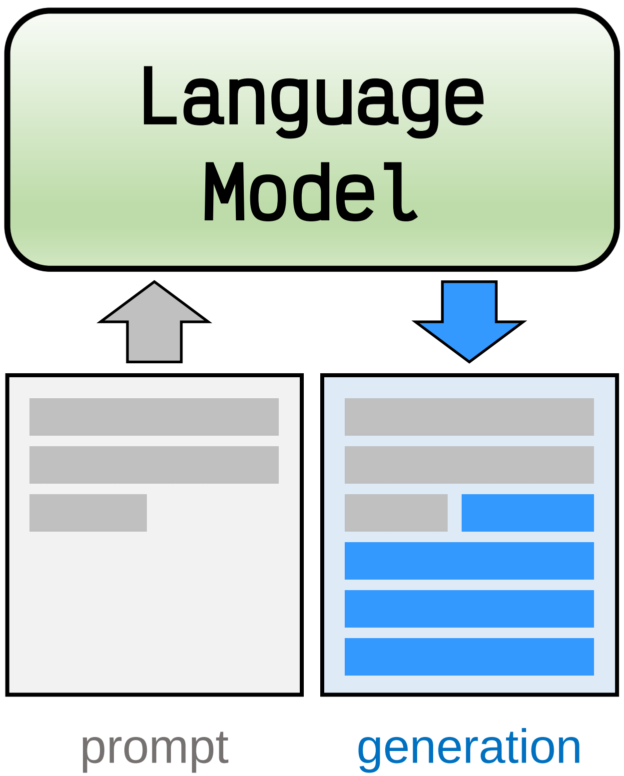


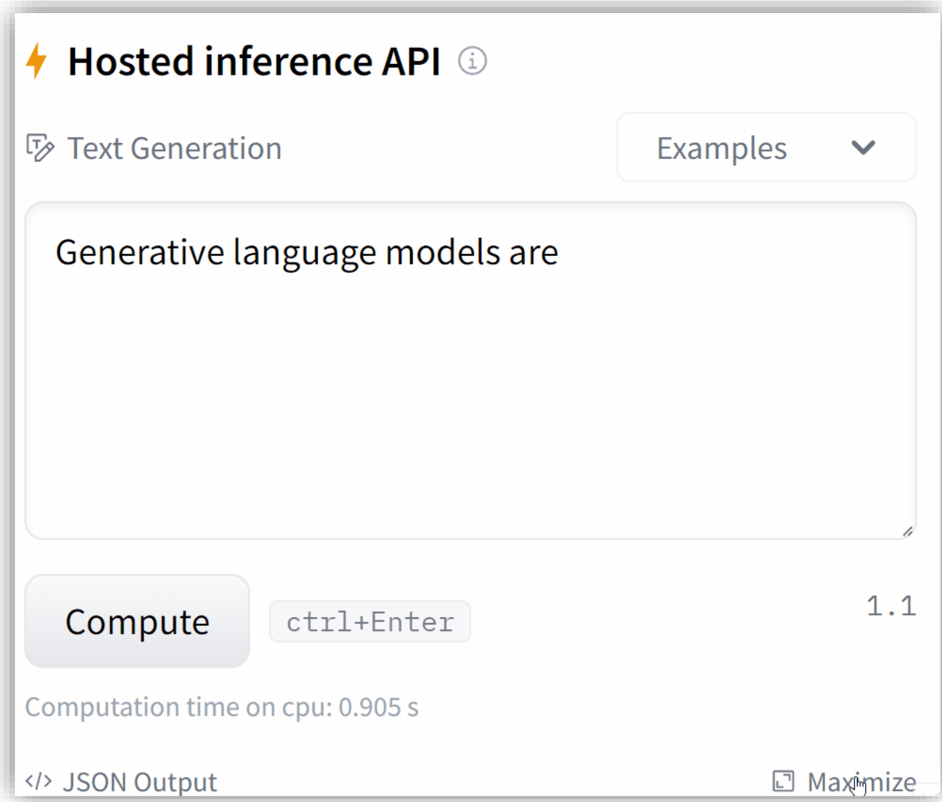
https://ai.googleblog.com/2022/03/auto-generated-summaries-in-google-docs.html

https://github.com/features/copilot

Generative Language Model

- Generate text that starts with prompt
  - E.g., GPT-2 [Radford et al. (2019)], GPT-3 [Brown et al. (2020)], …



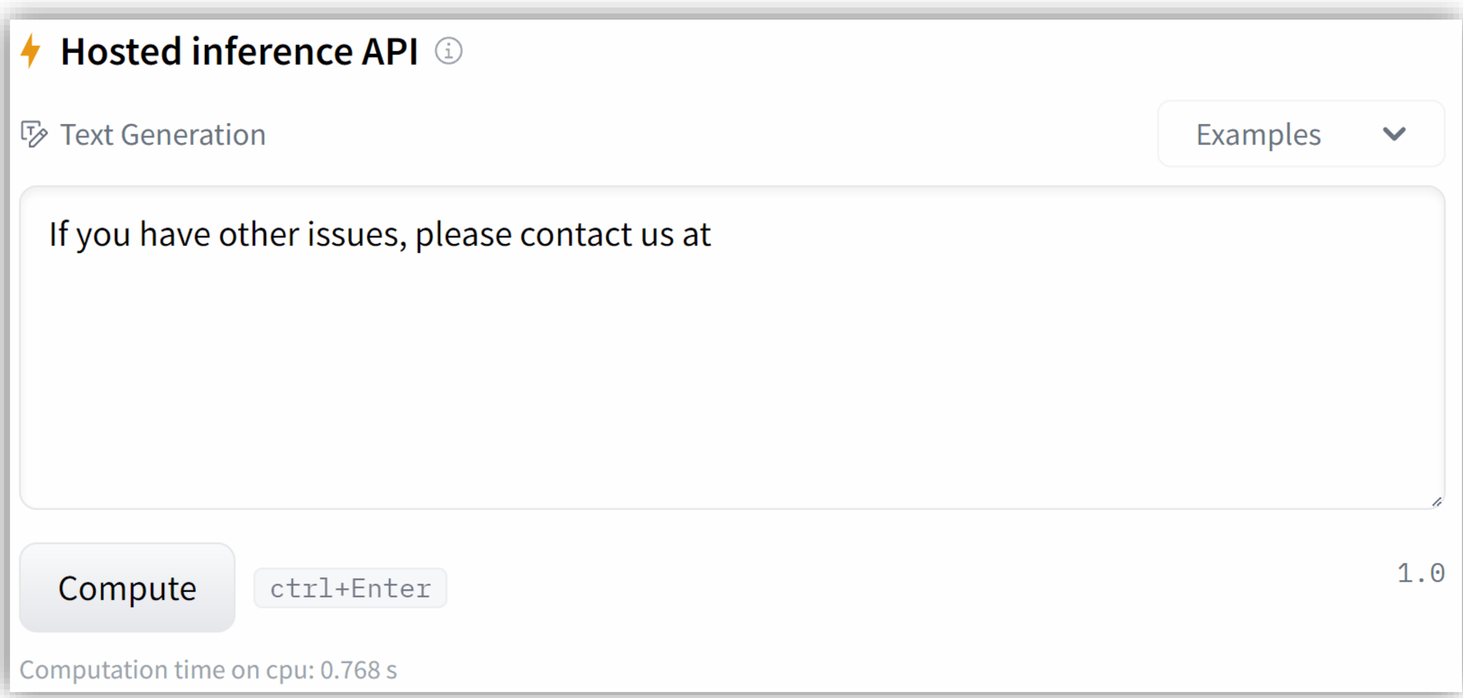


https://docs.cohere.ai/prompt-engineering-wiki, https://huggingface.co/gpt2

Unintentional Memorization in LM

- Generative LM can expose training data
  - Does the data real? (i.e., not synthesized?)

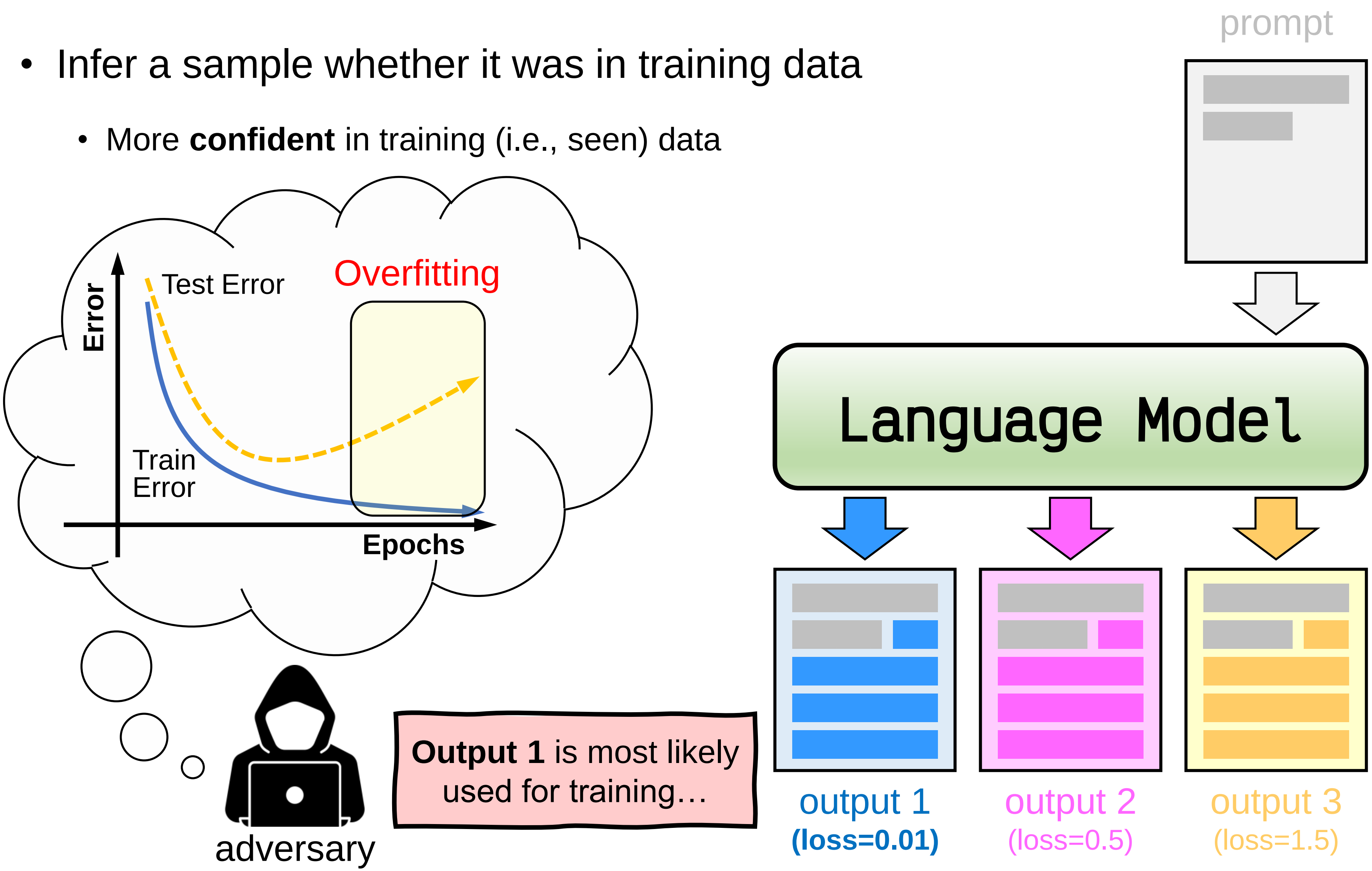
    → Determine by **membership inference** (MI) attack



https://huggingface.co/gpt2

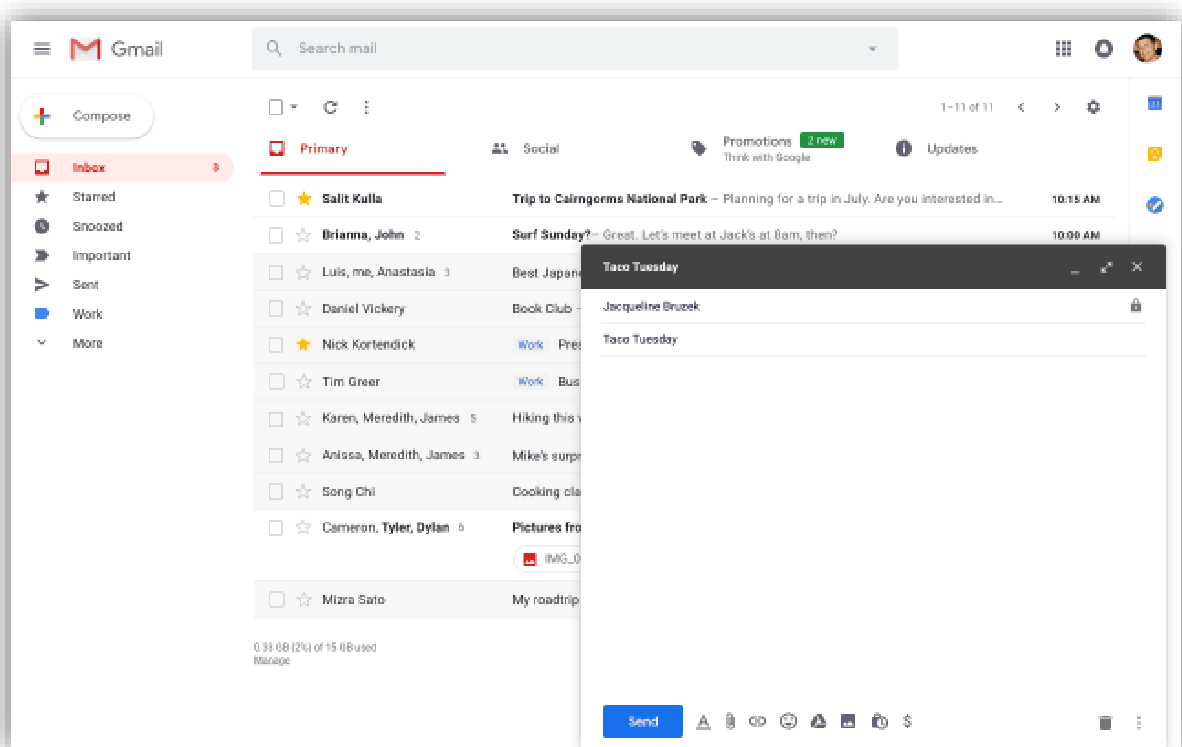
Membership Inference Attack

- Infer a sample whether it was in training data
  - More **confident** in training (i.e., seen) data



Membership Inference Attack

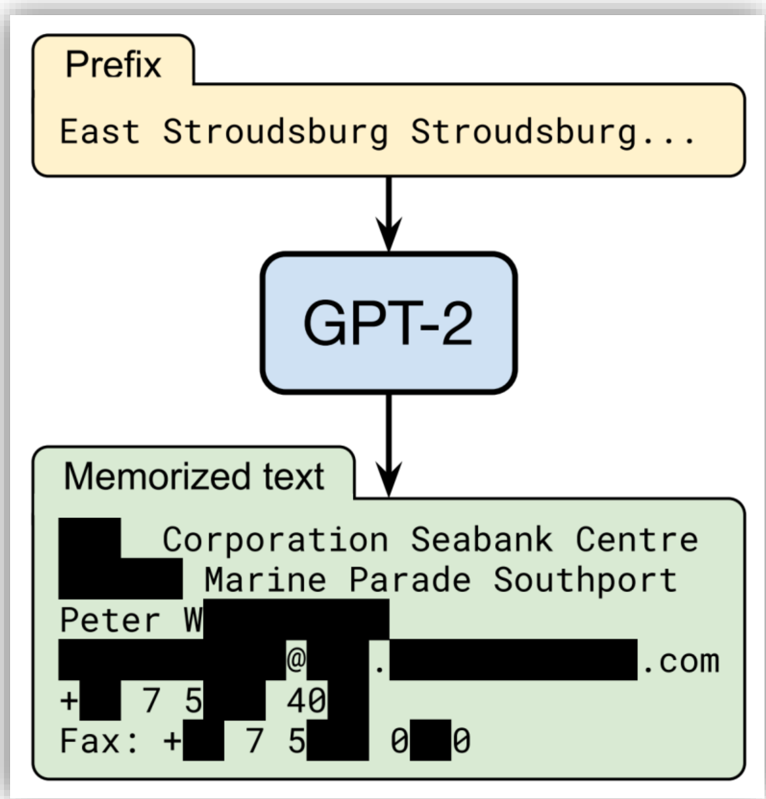
- Carlini et al. (2019, USENIX Security)
  - Quantitatively assessing the risk of unintentionally memorized
  - Google's Smart Compose (generative sequence model)



https://www.blog.google/products/gmail/subject-write-emails-faster-smart-compose-gmail/

Membership Inference Attack

- Carlini et al. (2021, USENIX Security)
  - Training data extraction attack on GPT-2 [Radford et al. (2019)]
  - Confirmed 604 training text sequences among 1,800 candidates



Observation: English-based LMs

- Previous works targeted **English-based LMs**
  - Carlini et al. (2019, USENIX Security)
  - **Carlini et al.** (2021, USENIX Security)
  - Carlini et al. (2022, IEEE S&P)
  - Carlini et al. (2022, arXiv)
  - Lee et al. (2022, ACL)
  - Lee et al. (2022, arXiv)
  - …

We raise a fundamental question,

- Are prior attacks **still effective** on **other language-based LMs**?
  - Spanish
  - Danish
  - Chinese
  - Japanese
  - **Korean**
  - …

Our Example: Korean

- Grammatical differences: **English** vs. **Korean**

	English	**Korean**
Spacing rules	Easy	Hard / Complex
Case-sensitive	True	False
Word orders	Strict	Flexible

I	go	to school	by	bus

↕

나는	버스를	타고	학교에	갑니다

Our Example: Korean

- Grammatical differences: **English** vs. **Korean**

	English	**Korean**
Spacing rules	Easy	Hard / Complex
Case-sensitive	True	False
Word orders	Strict	Flexible

I	go	to school	by	bus

↕

나는	버스를	타고	학교에	갑니다
나는	학교에	갑니다	버스를	타고
나는	학교에	버스를	타고	갑니다
나는	갑니다	학교에	버스를	타고
(나는)	버스를	타고	학교에	갑니다
학교에	버스를	타고	나는	갑니다
학교에	나는	버스를	타고	갑니다
학교에	갑니다	나는	버스를	타고
버스를	타고	나는	학교에	갑니다
버스를	타고	학교에	나는	갑니다
버스를	타고	갑니다	나는	학교에

⋮

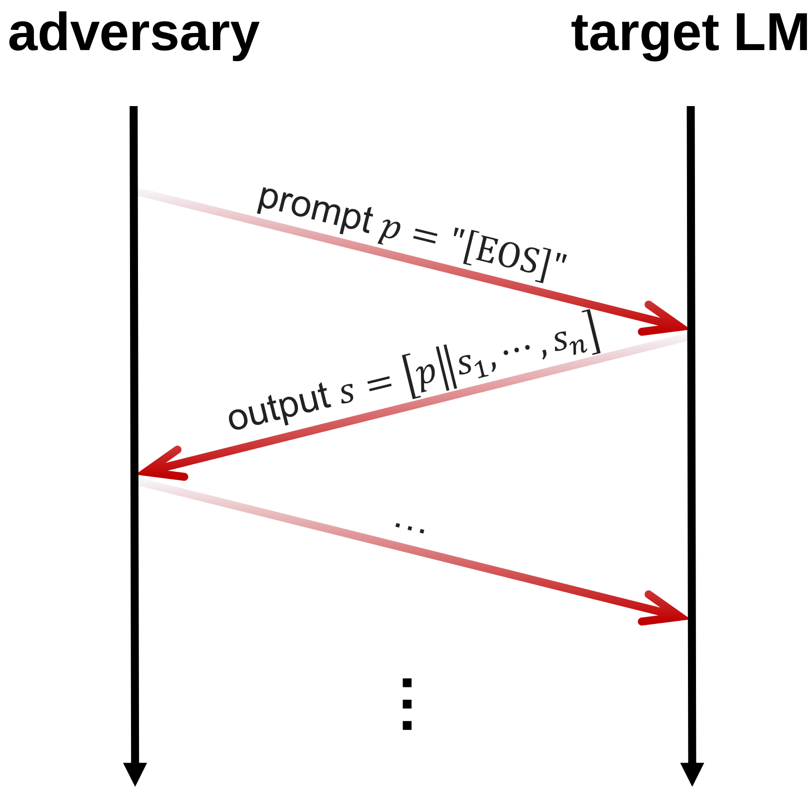
Methodology of this paper

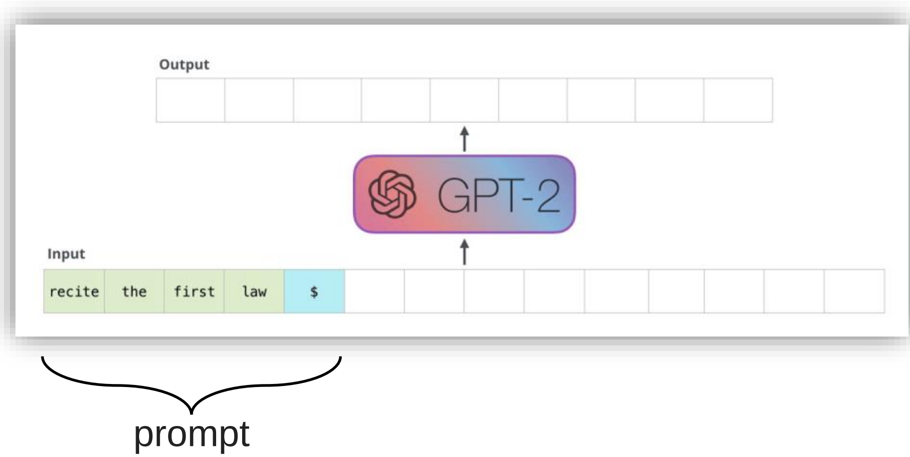
- Exploring the **effectiveness** and **uniqueness** of MI attack
  - **Effectiveness**: Perform Carlini's MI attack on KoGPT [Kim et al. (2021)]
  - **Uniqueness**: Increase the amount of information in top-k results

- Approach
  ① Step 1: **Text Sampling**
  ② Step 2: **Membership Inference**
  ③ Step 3: **Verification**

Step 1: Text Sampling

- Sample a sufficiently large number of texts
  - 100,000 texts & 256 tokens (w/o prompt)





https://jalammar.github.io/illustrated-gpt2/

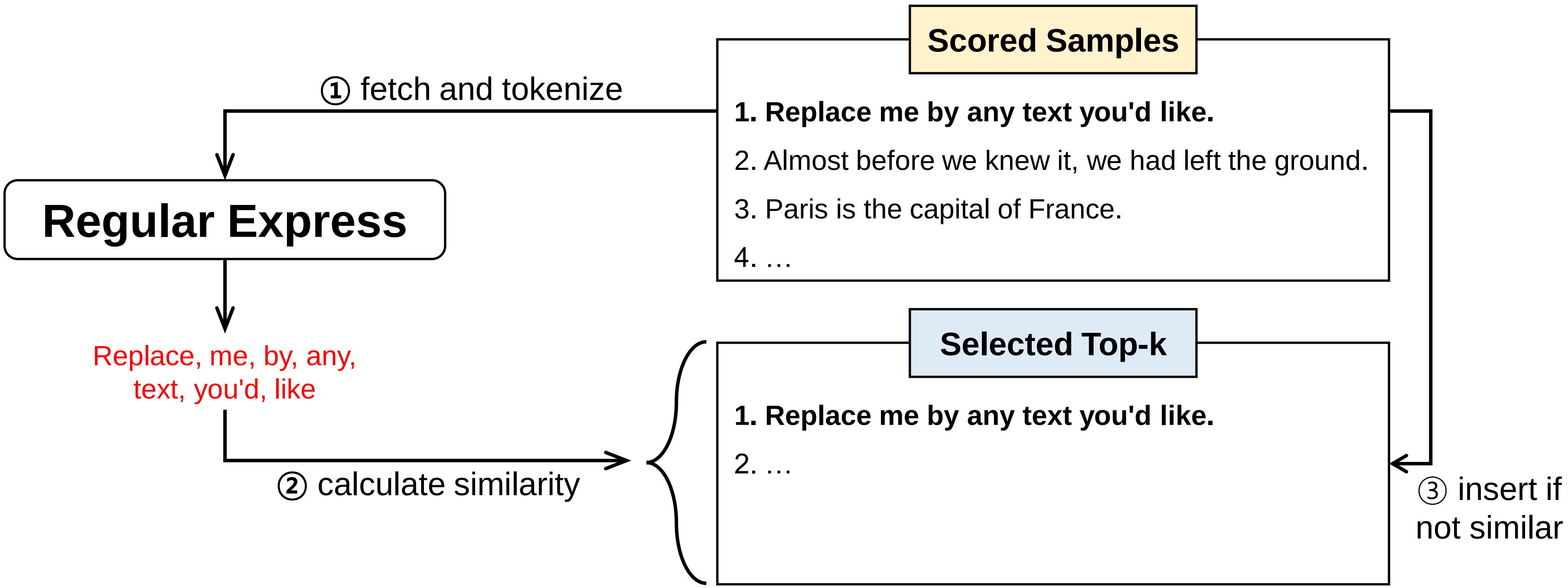
Step 2: Membership Inference

- Score samples
  - **Perplexity** (`PPL`) → How well does the LM predict?
  - **zlib entropy** (`zlib`) → How not repeated?
  - **Lowercase** (`Lowercase`) → How strange is the clean version?
  - **Sliding window** (`Window`) → How well does the LM predict, partially?

Shortcut	**Abstractive Evaluation Method**	**Lower Best**	**Higher Best**
`PPL`	$\log(\text{Perplexity})$	✓	-
`zlib`	(zlib Entropy) / `PPL`	-	✓
`Lowercase`	$\log(\text{Lowercase Perplexity})$ / `PPL`	-	✓
`Window`	$\log(\min\{\text{Perplexity of Sliding Windows}\})$	✓	-

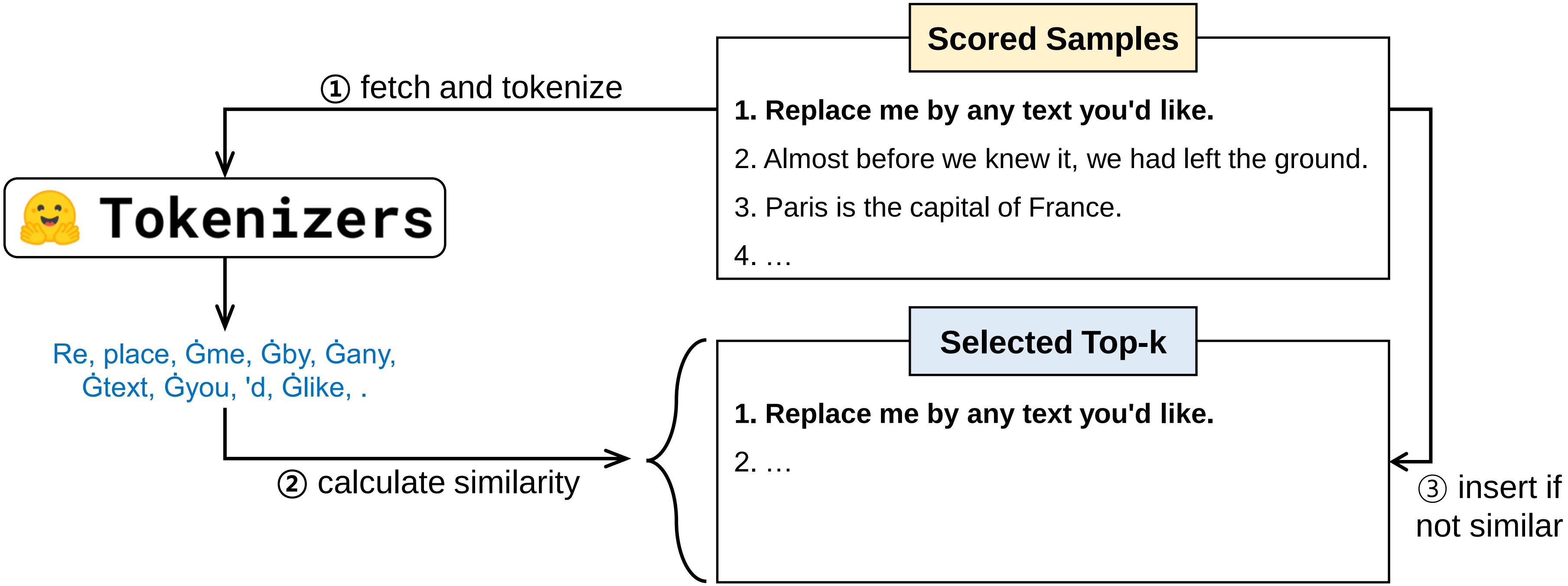
Step 3: Verification

- Select top-k samples per metric
  - Tokenize (word-level / Byte-Pair Encoding) ⇒ discerning enough?
  - Calculate trigram similarity
  - Choose sequentially not to overlap



Step 3: Verification

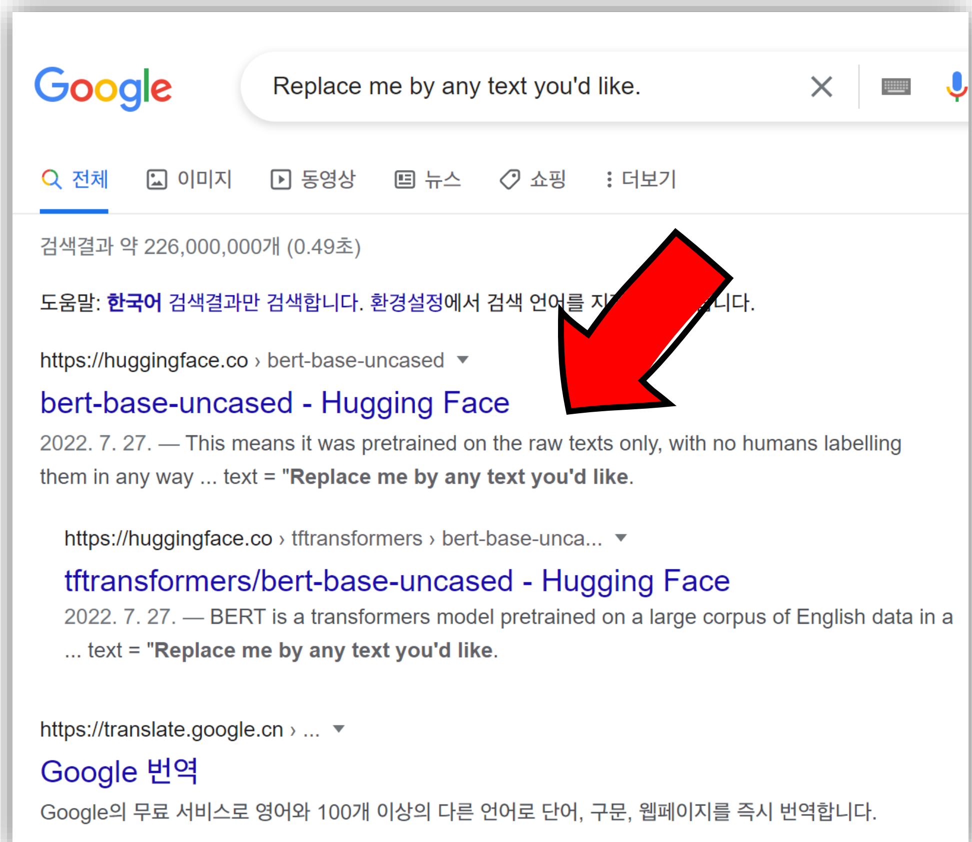
- Select top-k samples per metric
  - Tokenize (word-level / Byte-Pair Encoding) ⇒ discerning enough!
  - Calculate trigram similarity
  - Choose sequentially not to overlap



Step 3: Verification

- Select top-k samples per metric
  - Tokenize (word-level / Byte-Pair Encoding)
  - Calculate trigram similarity
  - Choose sequentially not to overlap

- Verify whether members or not
  - Manually search (Google)
  - Calculate (approximated) precision



Experiment Setup

- Environment
  - GPU: NVIDIA Quadro RTX 6000 (24GB) × 2
  - RAM: 157GB
  - DL framework: torch 1.11.0, transformers 4.17.0
- Target Model
  - KoGPT (Korean Generative Pre-trained Transformer) [Kim et al. (2021)]
  - 6.2B parameters

Research Questions

- (RQ1) **Effectiveness**
  - Still effective across language domains?
- (RQ2) **Uniqueness**
  - Does our proposal improve uniqueness?

Research Questions

- (RQ1) **Effectiveness**
  - Still effective across language domains?

- (RQ2) **Uniqueness**
  - Does our proposal improve uniqueness?

Evaluation 1: Effectiveness

- Remark (RQ1)
  - *Still effective across language domains?*

- Found with precision of 20% to 90%
  - `PPL`: Higher than expected (9% → 89%)
  - `Lowercase`: Low as expected (53% → 20%)

Target System	Metrics			
	`PPL`	`zlib`	`Lowercase`	`Window`
GPT-2 (XL)	9	59	**53**	33
KoGPT	**89**	**90**	20	**52**
Difference	↑ 80	↑ 31	↓ 33	↑ 19

Evaluation 1: Effectiveness

- Remark (RQ1)
  - *Still effective across language domains?*

- Found with precision of 20% to 90%
  - `PPL`: Higher than expected (9% → 89%)
  - `Lowercase`: Low as expected (53% → 20%)

Target System	Metrics PPL	zlib	Lowercase	Window
GPT-2 (XL)	9	59	**53**	33
KoGPT	**89**	**90**	20	**52**
Difference	↑ 80	↑ 31	↓ 33	↑ 19

Evaluation 1: Effectiveness

- Remark (RQ1)
  - *Still effective across language domains?* → **Yes, still effective**

- Found with precision of 20% to 90%
  - `PPL`: Higher than expected (9% → 89%)
  - `Lowercase`: Low as expected (53% → 20%)

Target System	Metrics			
	`PPL`	`zlib`	`Lowercase`	`Window`
GPT-2 (XL)	9	59	**53**	33
KoGPT	**89**	**90**	20	**52**
Difference	↑ 80	↑ 31	↓ 33	↑ 19

Examples of Inference Results

- Memorized Examples
  - **Bible verse**: Quoted as-is and undistorted
  - **Commercial stationery**: Repeated identical phrase
- Unmemorized Examples
  - **Wikipedia**: Matching only one line & HTML tags

Examples of Inference Results

- Memorized Examples
  - **Bible verse**: Quoted as-is and undistorted
  - **Commercial stationery**: Repeated identical phrase
- Unmemorized Examples
  - **Wikipedia**: Matching only one line & HTML tags

2015–16 Georgia State Panthers men's basketball team

From Wikipedia, the free encyclopedia

The 2015-16 Georgia State Panthers men's basketball team represented Georgia State University during the 2015-16 NCAA Division I men's basketball season. The team's head coach was Ron Hunter in his fifth season. The Panthers played their home games at the GSU Sports Arena and competed as a member of the Sun Belt Conference. They finished the season 16-14, 9-11 in Sun Belt play to finish in sixth place. They lost in the first round of the Sun Belt Tournament to Texas State

Contents [hide]

2015-16 Georgia State Panthers men's basketball

GEORGIA STATE

Conference	Sun Belt Conference
Record	16-14 (9-11 Sun Belt)
Head coach	Ron Hunter (5th season)
Assistant coaches	Darryl LaBarrie (5th season) Everick Sullivan (5th season) Claude Pardue (5th season)
Home arena	GSU Sports Arena

Seasons

← 2014-15 2016-17 →

Lowercase Top-16

The 2015–16 Georgia State Panthers men's basketball team represented Georgia State University during the 2015–16 NCAA Division I men's basketball season. The Panthers were coached by Bill Self and played their home games at the GSU Sports Arena. They were members of the Atlantic Coast Conference. They finished the season 14–17, 8–16 in ACC play to finish in twelfth place. They advanced to the semifinals of the Atlantic Coast Tournament where they lost to NC State.
Schedule.
!colspan=9 style="background:#002299; color:#FFFFFF;"| Exhibition
!colspan=9 style="background:#002299; color:#FFFFFF;"| Non-conference regular season
!colspan=9 style="background:#002299; color:#FFFFFF;"| ACC regular season
!colspan=9 style="background:#002299; color:#FFFFFF;"| Atlantic Coast Tournament
!colspan=9 style="background:#002299; color:#FFFFFF;"| National Invitation Tournament
Rankings.
2015–16 NCAA Division I Men's Basketball season
NCAA Division

Evaluation 2: Uniqueness

- Remark (RQ2)
  - *Does our proposal improve uniqueness?*

- Found 6% to 22% of underestimates
  - `Window`: Higher than expected (22%)

Eval. Item	**Metrics**			
	`PPL`	`zlib`	`Lowercase`	`Window`
# of Underrated	6	6	7	**22**

Evaluation 2: Uniqueness

- Remark (RQ2)
  - *Does our proposal improve uniqueness?* → **Yes, improve uniqueness**

- Found 6% to 22% of underestimates
  - `Window`: Higher than expected (22%)

Eval. Item	Metrics			
	`PPL`	`zlib`	`Lowercase`	`Window`
# of Underrated	6	6	7	**22**

Evaluation 2: Uniqueness

- Remark (RQ2)
  - *Does our proposal improve uniqueness?* → **Yes, improve uniqueness**

- (Future work) Is increasing uniqueness **efficient**?
  - `Lowercase`: 2%p ↓
  - The others: 1%p to 7%p ↑
  - Overall, increasing uniqueness **increases precision**

Target System	Tokenization	Metrics			
		`PPL`	`zlib`	`Lowercase`	`Window`
GPT-2 (XL)	Word-level	9	59	**53**	33
KoGPT	Word-level	89	90	20	52
KoGPT	BPE	**91**	**91**	18	**59**

Is the extracted data sensitive?

- Most were preprocessed,
  - Message *** **** ****
  - Account number: KEB ***-**-*****-*
  - …

Is the extracted data sensitive?

- Most were preprocessed,
  - Message *** **** ****
  - Account number: KEB ***-**-*****-*
  - …
- But some were not
  - Phone number
  - Code (HTML, …)
  - …
- (cf.) It may not be privacy
  - Corporation or organization

How to mitigate the leakage?

- Differential privacy (DP) training [Dwork et al. (2006), Dwork (2008)]
  - Guarantee the privacy of training data
  - Tradeoff exists: **preserving privacy** vs. **utility** (e.g., accuracy)

- Deduplicating training data [Lee et al. (2022)]
  - Reduce memorization & increase generalization

    → Reduce unintentional leakage
  - Save training time

How to mitigate the leakage?

- Differential privacy (DP) training [Dwork et al. (2006), Dwork (2008)]
  - Guarantee the privacy of training data
  - Tradeoff exists: **preserving privacy** vs. **utility** (e.g., accuracy)

- Deduplicating training data [Lee et al. (2022)]
  - Reduce memorization & increase generalization

    → Reduce unintentional leakage
  - Save training time

Conclusion

- **Confirm the effectiveness**
  - Prior works still effective across language domains; 20% to 90% precision

- **Improve the uniqueness**
  - Increase the amount of information contained; 6% to 22% in top-k

Thank you!

Code: https://github.com/seclab-yonsei/mia-ko-lm

Email: myunggyo.oh@yonsei.ac.kr (Myung Gyo Oh)

More about future work

- Language Model Memorization
  - How do we **define** the language model memorization?
  - **Why** does the language model memorize?
  - How do we **mitigate** the memorization?
  - …
- Extension to memorization of the deep learning models
  - Reduce memorization → improve generalization

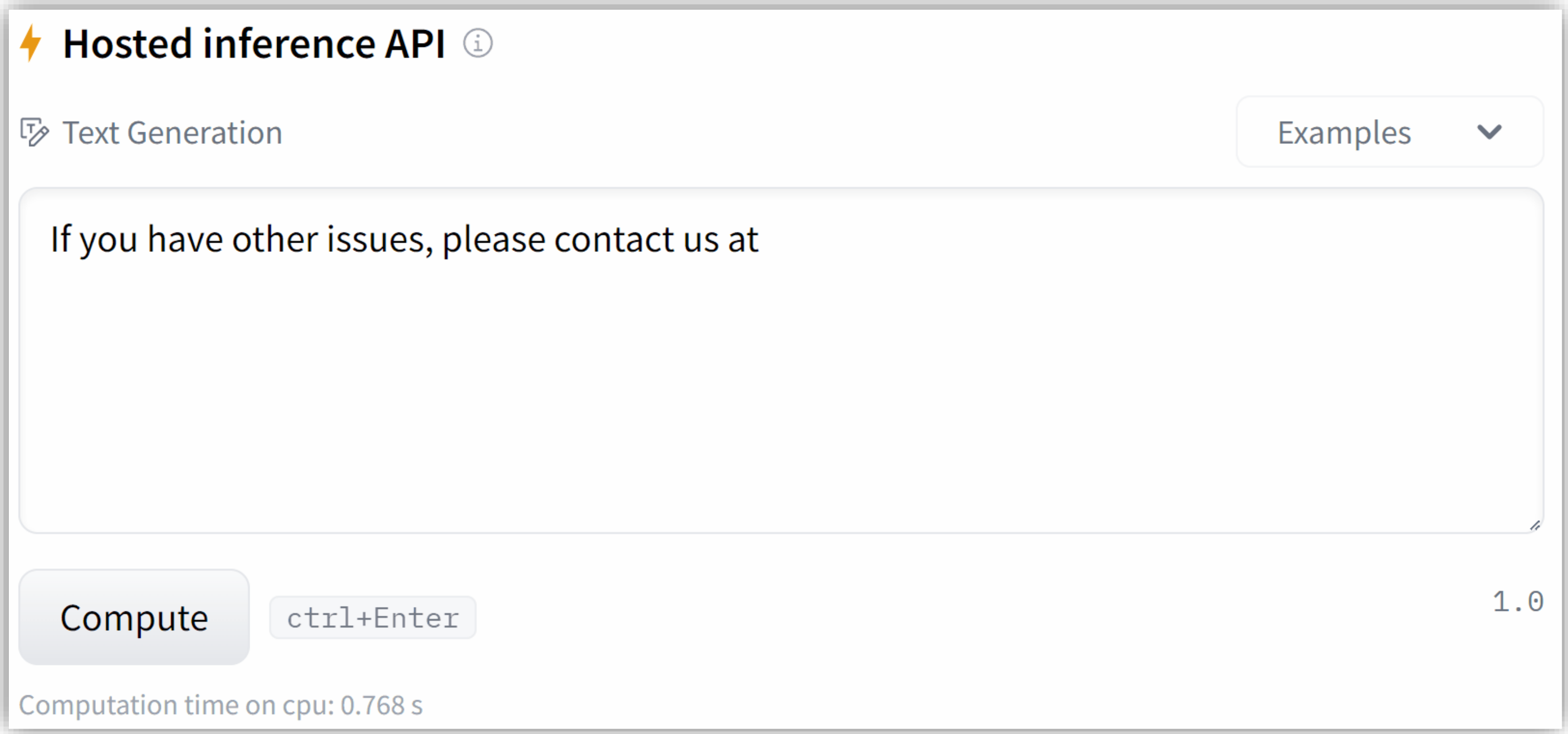
Why not use shadow training?

- **Limitation 1: Training Strategy** (slight issue)
  - Shadow training: *Supervised learning*
  - Recent language models: *Unsupervised learning*

- **Limitation 2: High Cost** (significant issue)
  - Scale of recent LMs and datasets rises exponentially (e.g., GPT-3 [Brown et al., 2020] → 175 billion parameters)
  - So, it's hard to { generate | train | infer } multiple shadow models

Does this really happens?

- Unfortunately, yes, it does 😥
  - **Targeted attacks** can increase the likelihood of leaking specific data



Why is MI attack significant?

- **Breaking** Confidentiality of Training Data

- **Leaking** Personal Information
  - Exploits the property: Appeared in fewer documents → more sensitive [Carlini et al., 2021]

- **Infringing** Copyright or Intellectual Property Rights
  - Training data exposure can lead to **plagiarism** (e.g. newspaper, article, …)

Advice to apply in other languages

- Check the baseline attacks defined on English-based LMs
  - English is the **official language** → **still** (somewhat) **effective** in other languages

- Modify the attacks specific to language domains you want

> 아래 동영상들을 클릭하세요 01. Consuelo"s Love Theme / James Galway & Cleo Laine 영국 출신의 백인여성 재즈가수로 1980년 작품. 02. Jeg Ser Deg Sote Lam (당신곁에 소중한 사람) / Susanne Lundeng 스웨덴 출신의 월드 뮤직 연주자로 1997년 작품. 03. Calcutta / Lawrence Welk 이지리스링 연주 악단 04. Amsterdam Sur Eau (물위의 암스테르담) / Claude Ciari 프랑스 출신의 팝 기타리스트"끌로드 치아리"의 70년대 말 작품 으로 멋과 낭만이 깃든 감미로운 연주곡. 끌로드 치아리는 63년 첫 솔로작"HUSHABYE"를 발표 한 후 일약 스타로 뛰어 오른 팝 기타리스트로 주요 작품으로는"첫 발자욱"과 함께 "LA PLAYA","US$$","Soul Of A Man"등이 있다. 05. Recuerdos De La Alhambra (알함브라 궁전의 추억) / Narciso Yepes 1927년 스페인 동남부의 로르카니 출신의 작곡가 겸 기타리스트. 1952년 프랑스 영화(금지된 장난)의 음악을 맡

Curious Examples

- Language Model as **Archives**
  - Even extracted an article from 2002 (*estimated*)

불매운동은 계속 되야한다.쭈~욱

30774 이애숙 [annie] 2002-03-12

(1) 미국 불매운동, 왜 시작되었는가?

여기에 대해서 사람들마다 여러 가지 의견이 나오고 있습니다.

첫 번째로 나오는 이야기가, "미국 제품 불매운동을 왜 하여야 하는가?"라는 것입니다.

그 이유는 약 3주 전으로 거슬러 올라갑니다.

우리나라 김동성 선수는 쇼트트랙에서 금메달을 차지하고도, 비열한 아폴로 안톤 오노와, 치사한 미국 심판들의 짝짜꿍으로 금메달을 빼앗기게 됩니다.

우리나라 사람들은 분노했습니다.

1. 미국 불매운동,왜 시작되었는가? 여기에 대해서 사람들마다 여러 가지 의견이 나오고 있습니다. 첫 번째로 나오는 이야기가, "미국 제품 불매운동을 왜 하여야 하는가? "라는 것입니다. 그 이유는 약 3주 전으로 거슬러 올라갑니다. 우리나라 김동성 선수는 쇼트트랙에서 금메달을 차지하고도, 비열한 아폴로 안톤 오노와, 치사한 미국 심판들의 짝짜꿍으로 금메달을 빼앗기게 됩니다. 우리나라 사람들은 분노했습니다. 하지만 미국은 우리나라에게만 편파판정을 내린 것이 아니었습니다. 가장 피해를 많이 입은 것은 "러시아, 일본, 한국"이 세 나라로 좁혀집니다. 당연히 세 나라 국민들은 강력하게 반발했고, IOC에 항의도 하였습니다. 그 결과, 러시아와 일본에게는 "편파판정 사과"를, 우리나라의 항의는 기각되었죠. (가장 피해를 입은 것은 김동성 선수인데 말입니다.) 그리고, 그것만으로 미국은 끝내지 않았습니다. 미국 방송들의 한국 비하와, 인종 차별적 발언을 서슴치 않았고, 더러운 안톤 오노를 영웅으로 추켜세웠습니다. 우리나라 사람

https://bbs.catholic.or.kr/homem/bbs_view.asp?num=100861&id=142495&menu=bbs1plaza

Curious Examples

- **Window** chooses **uninteresting samples** in top-k
  - E.g., repeated hyphens, ...

이 사고로 이씨와 조씨 및 김모(57·여)씨 등 3명이 중상을 입어 인근 병원으로 옮겨졌으나 김씨는 결국 숨져 주위를 안타깝게 하고 있다. 이씨와 조씨는 머리를 크게 다쳐 의식불명 상태로 병원에서 치료 중이나 생명도 위태로운 것으로 알려졌다. 경찰은 이들이 불법으로 영업 중인 음식점에 손님으로 가장해 들어간 뒤 나오던 중 식당 밖에 서 있던 다른 손님을 치기위해 후진을 하던 중 중심을 잃고 쓰러지자 이를 피하려다 사고가 난 것으로 보고 정확한 경위를 조사중이다.(끝) 〈 긴급속보 SMS 신청 〉〈 포토 매거진 〉〈 M-SPORTS 〉

(런던.AFP=연합) 최근 발생한 런던 폭탄 테러가 영국 국민의 단결을 요구한 보리스 판크스 前내무장관에 대한 보복인 것으로 믿어지고 있다고 집권 보수당의 한 당직자가 30일 밝혔다. 이 당직자는 이날 BBC와의 회견에서 "그들은 보리스 판크스가 아직도 우리나라를 통제하고 있다고 믿으며 그를 제거해야 한다는 것을 강조하고 있다"며 이같이 말했다. 그는 또 런던 테러 이후 실시된 긴급 여론조사 결과 대부분의 사람들이 범인들이 "비인간적인 동기를 가진 범죄자들로서 영국에 대해 증오심을 갖고 있으며 그들의 행위는 국가를 위해한 것으로 봐야 한다"는 데 동의한 것으로 나타났다고 덧붙였다.(끝)

★ ★세계는 지금 ★

제가 쓰는 이 소설 속엔 사람들이 잘 생각하지 못하는 것들이 담겨 있는 듯이 생각됩니다. 예를 들어 제 소설 속의 인물은 그들의 욕망과 그들의 감정을 드러내기 위해, 그들의 본능을 대변하기 위해, 다른 한편 제 소설 속의 인물은 그들의 욕망을 반영하기 위해, 그들의 본능을 반영하기 위해. 제 소설 속의 '당신'은 그들의 존재를 투영해 내기 위해, 그리고 또 다른 소설 속의 '당신'은 그들의 존재를 투영해 내기 위해, 그리고 또 다른 소설 속의 '당신'은 그들의 존재를 나타내기 위해. 제 소설 속의 '나'는 당신을 투영해 내기 위해, 그리고 다른 소설 속의 '나'는 당신의 존재를 투영해 내기 위해. - [나는 나를 파괴할 권리가 있다] 중에서....... ★
공지어겼다면.. 메일(***************@*******.***) ★연재소설